DOCUMENT RESUME

ED 386 492 TM 024 061

AUTHOR	Livingston, Samuel A.; Lewis, Charles
TITLE	Estimating the Consistency and Accuracy of Classifications Based on Test Scores.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-93-48
PUB DATE	Oct 93
NOTE	31p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Classification; Error of Measurement; *Estimation (Mathematics); *Reliability; *Scores; Scoring; Test Length; Test Results; *Test Use
IDENTIFIERS	Accuracy; Beta Binomial Test Model; Four Parameter Model

ABSTRACT

This paper presents a method for estimating the accuracy and consistency of classifications based on test scores. The scores can be produced by any scoring method, including the formation of a weighted composite. The estimates use data from a single form. The reliability of the score is used to estimate its effective test length in terms of discrete items. The true-score distribution is estimated by fitting a four-parameter beta model. The conditional distribution of scores on an alternate form, given the true score, is estimated from a binomial distribution based on the estimated effective test length. The agreement between classifications on two alternate forms is estimated by assuming conditional independence, given the true score. An evaluation of the method showed that the estimates of the percent of test-takers correctly classified and the percent consistently classified were within one percentage point of the actual values in most cases. Although the estimated effective test length and the estimates of the conditional standard error of measurement are sensitive to changes in the specified minimum and maximum possible scores, the estimates of the decision accuracy and decision consistency statistics are not. (Contains seven tables and seven references.) (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

RR-93-48

ED 386 492

RESEARCH REPORT

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
H. I. BRAUN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ESTIMATING THE CONSISTENCY AND ACCURACY OF CLASSIFICATIONS BASED ON TEST SCORES

Samuel A. Livingston
Charles Lewis

BEST COPY AVAILABLE

Educational Testing Service
Princeton, New Jersey
October 1993

TM024061

Estimating the Consistency and Accuracy of Classifications Based on Test Scores

Samuel A. Livingston
Charles Lewis

Educational Testing Service

Copyright © 1993. Educational Testing Service. All rights reserved.

Abstract

This paper presents a method for estimating the accuracy and consistency of classifications based on test scores. The scores can be produced by any scoring method, including the formation of a weighted composite. The estimates use data from a single form. The reliability of the score is used to estimate its effective test length in terms of discrete items. The true-score distribution is estimated by fitting a four-parameter beta model. The conditional distribution of scores on an alternate form, given the true score, is estimated from a binomial distribution based on the estimated effective test length. The agreement between classifications on two alternate forms is estimated by assuming conditional independence, given the true score.

An evaluation of the method showed that the estimates of the percent of test-takers correctly classified and the percent consistently classified were within one percentage point of the actual values in most cases. Although the estimated effective test length and the estimates of the conditional standard error of measurement are sensitive to changes in the specified minimum and maximum possible scores, the estimates of the decision accuracy and decision consistency statistics are not.

Acknowledgments

The project described in this paper is the work of a group of Educational Testing Service staff members. Ruth Mroczka and Marilyn Wingersky wrote the computer program. The true-score estimation procedure is based on the work of Frederic Lord. Carole Bleistein prepared and analyzed the data for tryouts of the method and its previous versions.

Estimating the Consistency and Accuracy of Classifications Based on Test Scores

Samuel A. Livingston
Charles Lewis

The problem

Several authors have proposed methods for estimating the accuracy or consistency of classifications based on test scores (e.g., Huynh, 1976; Subkoviak, 1976; Livingston and Wingersky, 1979; Wilcox, 1981). All of these methods are based on the assumption that the test consists of a known number of equally weighted items, scored simply as correct or incorrect, and that the test score is the number of those items answered correctly. This situation is certainly a common one. However, many tests are not scored in this way. Essay tests and performance assessments typically are scored in a way that allows for partial credit on each item. In some testing programs, test-takers are classified on the basis of a composite score -- a weighted sum of scores on two or more tests or subtests, which may be unequally weighted. In all of these cases, determining the effective length of the test used as the basis for classification is not just a simple matter of counting test items.

The purpose of this paper is to suggest a generally applicable method for using data from one form of a test to estimate the accuracy and the consistency of classifications based on the scores. This method applies not only to test scores determined by counting correct answers, but to any test score for which a reliability coefficient can be estimated. The method described here is actually the fourth in a series of solutions to this problem. As we have progressed from the first to the fourth solution, the method has become simpler, more generally applicable, and more accurate.

Terminology

In this paper we will use the term "test" to refer to the entire measurement procedure that produces the score that is used as the basis for classification. The "test", as we will use the term, could actually be a battery of several measures, possibly including such measures as interview-based evaluations and performance ratings as well as paper-and-pencil tests. We will use the term "test forms" to refer to independent replications of the measurement procedure, varying all the factors that are to be considered as contributing to errors of measurement. These factors could include the specific questions or problems presented to the test-taker, the raters or scorers of any subjective portions, the examiners or interviewers (to the extent that they may affect the outcome of the measurement), and possibly other factors also.

We will not use the term "observed score" for the variable that describes a test-taker's score on a single form of the test, because in some cases we will be referring to a score on a hypothetical alternate form of the test, which is not actually observed. Therefore, we will call this variable a "single-form score". We will use the term "true score" as the term is traditionally used in psychometrics: an expected (average) value of the test score, averaged over those factors classified as measurement error. We will use the term "true score" because it is concise and familiar; a more precise and more descriptive term would be "all-forms average".

The term "accuracy", as used in this paper, refers to the extent to which the actual classifications of test-takers (on the basis of their single-form scores) agree with those that would be made on the basis of their true scores, if their true scores could somehow be known. The term "consistency"

refers to the agreement between the classifications based on two non-overlapping, equally difficult forms of the test. The group of test-takers for whom we want to estimate these accuracy and consistency statistics will be referred to as the "test-taker population".

We will use the term "effective length" to refer to a property of the test closely related to the precision of the scores. It is the number of discrete, dichotomously scored, locally independent, equally difficult test items necessary to produce total scores having the same precision as the scores being used to classify the test-takers. For example, the test might consist of three essay questions; its possible score range might be 30 points. But if it had the same reliability in the test-taker population as a test made up of 24 discrete items scored simply as right or wrong, its effective length would be 24.

The points that divide the score scale into categories for classifying the test-takers will be referred to as "cut-points".

Input

The method presented here requires four kinds of information as input: (1) the distribution of the scores on one form of the test, observed or estimated for the test-taker population, (2) the reliability coefficient of the scores, computed or estimated for the test-taker population, (3) the maximum and minimum possible scores on the test, and (4) the cut-points that separate the categories. Although the test score may take on a very large number of possible values -- enough to make it effectively a continuous variable -- the computational procedure requires the scores to be expressed as integers. The cut-points are assumed to be halfway between the highest score in one category and the lowest score in the next.

Output

Tables 1, 2, 3, and 4 contain examples of some of the statistics produced by the method described in this paper. They are based on a case in which the test-takers were classified into five categories by applying four cut-points. Tables 1 and 2 contain statistics describing decision accuracy -- the agreement between the classifications based on the form actually taken and the classifications that would be made on the basis of the test-takers' true scores. Table 1 is the 5x5 contingency table that results from applying all four cut-points simultaneously. Table 2 consists of the four separate 2x2 contingency tables that result from applying the four cut-points separately.

Tables 3 and 4 correspond to Tables 1 and 2, except that they contain statistics describing decision consistency -- the agreement between the classifications based on the form actually taken and the classifications that would be made on the basis of an alternate form. Notice that the agreement indicated by Tables 3 and 4 is not as strong as the agreement indicated by Tables 1 and 2. The reason is that in Tables 3 and 4, each variable includes a random component. In Tables 1 and 2, only one of the variables includes a random component.

Tables 1-4 show some particularly useful statistics for describing decision consistency and decision accuracy. However, there are other useful statistics that are not shown in these tables. In particular, the test user might be interested in the conditional distribution of single-form scores for test-takers with a given true score, or in the standard deviation of this conditional distribution, which is the conditional standard error of measurement. The method we describe estimates these statistics also.

Notation

In this paper, X will represent a test-taker's score on one form of the test, rounded to the nearest integer. Where necessary, we will identify scores on alternate forms of the test by X_0, X_1, and X_2; X_0 will represent the form of the test for which data are available. The lowest and highest possible scores will be represented by X_{min} and X_{max}. (X_{min} may be negative.)

The reliability coefficient of the test scores in the test-taker population will be represented by **r**. The estimated effective length of the test will be represented by **n**.

Sometimes it will be convenient to express the single-form score on a scale of 0 to 1. This proportional score will be represented by **p**, so that

$$p = \frac{X - X_{min}}{X_{max} - X_{min}} . \qquad (1)$$

The "true score" associated with X will also be expressed as a proportion, on a scale of 0 to 1, and represented by T_p, so that

$$T_p = \frac{E(X) - X_{min}}{X_{max} - X_{min}} . \qquad (2)$$

We will refer to T_p as the "proportional true score".

Some parts of the estimation procedure involve a rescaling of the single-form score from its original scale, which extends from X_{min} to X_{max}, onto a new scale that extends from 0 to **n**. This transformed score will be represented by X', so that

$$X' = np = n \frac{X - X_{min}}{X_{max} - X_{min}} . \qquad (3)$$

Table 5 summarizes these three score scales.

An overview of the method

The general procedure can be described as a series of steps:

1. Estimate the effective test length (n).

2. Estimate the distribution of the proportional true scores (T_p). Divide the range of this distribution into "levels" (intervals) of size .01 and compute the proportion of the distribution at each level of T_p.

3. At each level of T_p, construct a binomial distribution with parameters n and T_p. This is the distribution of scores on a hypothetical test of n discrete items, for a test-taker with proportional true score T_p. Call this score variable X'.

4. Transform the category boundaries linearly from the original scale of X (from X_{min} to X_{max}) onto the scale of X' (from 0 to n). Use these transformed boundaries to determine, for each level of T_p, the conditional classification on a single form of the test (X_1) other than the form actually taken.

5. Also transform the category boundaries linearly from the original scale of X (from X_{min} to X_{max}) onto the scale of T_p (from .00 to 1.00). Use these transformed boundaries and the conditional classifications from Step 4 to estimate the joint distribution of classifications on proportional true scores (T_p) and scores on a single form (X_1).

6. At each level of T_p, use the conditional classification on X_1 (from Step 4) to construct a conditional two-way classification on X_1 and X_2, where X_2 is the score on another form of the test. The assumption is that classifications based on X_1 and X_2 are independent and identically distributed for test-takers at a given level of T_p (i.e., a given interval of size .01). Then sum over the levels of T_p, to get the

estimated two-way classification based on X_1 and X_2 for the full test-taker population.

7. Adjust the estimated two-way classification based on T_p and X_1 (produced in Step 5) so that the category frequencies of X_1 will match those observed for X_0. The adjustment consists of determining a multiplier for each category of X and applying it to all the cell frequencies for that category of X. Use this adjusted two-way distribution of T_p and X_0 as the basis for estimating statistics that describe decision accuracy.

8. Adjust the estimated two-way classification on X_1 and X_2 (produced in Step 6) so that the category frequencies of X_1 will match those observed for X_0. Again, the adjustment consists of determining a multiplier for each category of X_1 and applying it to all the cell frequencies for that category of X_1. (There is no attempt to make the category frequencies of X_2 match those observed for X_0.) Use this adjusted two-way classification on X_0 and X_2 as the basis for estimating statistics that describe decision consistency.

The next two sections of this paper explain some of these steps in greater detail.

Estimating the effective test length

The effective test length corresponding to a test score is the number of discrete, dichotomously scored, locally independent, equally difficult items required to produce a total score of the same reliability. The effective test length of a score can be estimated from its mean, variance, and reliability coefficient in the test-taker population. If **r** is the reliability coefficient

of score X, **r** is also the reliability coefficient of the proportional score **p**. Similarly, if **n** is the effective test length of X, **n** is also the effective test length of **p**.

The key to estimating the effective length of **p** (and therefore of X) is to find two expressions for the overall error variance in **p** and set them equal. One of these expressions is based on the reliability coefficient:

$$\sigma_e^2 = \sigma_p^2(1-r) \quad . \tag{4}$$

The other expression for the overall error variance in **p** is found by observing that, at any given level of T_p, <u>all</u> the variance in **p** is error variance. The expectation of this conditional variance, over the population distribution of T_p, is the overall error variance of **p**:

$$\sigma_e^2 = E[Var(p|T_p)] \quad . \tag{5}$$

Setting these two expressions equal,

$$\sigma_p^2(1-r) = E[Var(p|T_p)] \quad . \tag{6}$$

At this point it is necessary to make an assumption: that the conditional error variance of scores on an **n**-item test, for test-takers with proportional true score T_p, is the variance of a binomial distribution based on **n** observations with success probability T_p. With this assumption, Equation 6 becomes

$$\sigma_p^2(1-r) = E\left[\frac{T_p(1-T_p)}{n}\right] = \frac{1}{n}\left[E(T_p) - E(T_p^2)\right] . \quad (7)$$

The expectation is over the distribution of proportional true scores (T_p) in the test-taker population. $E(T_p)$ is the mean proportional true score, which is equal to the mean proportional score on a single form. Call this mean score μ_p. Then

$$E(T_p) = \mu_p . \quad (8)$$

To find $E(T_p^2)$, note that for T_p as for any other variable,

$$Var(T_p) = E(T_p^2) - [E(T_p)]^2 , \quad (9)$$

so that

$$E(T_p^2) = Var(T_p) + [E(T_p)]^2 = r\sigma_p^2 + \mu_p^2 , \quad (10)$$

because **r**, the reliability coefficient, is the ratio of the variance of T_p to the variance of **p**.

Substituting these results into the right side of Equation 7,

$$\sigma_p^2(1-r) = \frac{1}{n}\left[\mu_p - (r\sigma_p^2 + \mu_p^2)\right] = \frac{1}{n}\left[\mu_p - \mu_p^2 - r\sigma_p^2\right] = \frac{1}{n}\left[\mu_p(1-\mu_p) - r\sigma_p^2\right] . \quad (11)$$

Solving for **n**, the effective test length,

BEST COPY AVAILABLE

$$n = \frac{\mu_p(1-\mu_p) - r\sigma_p^2}{\sigma_p^2(1-r)} . \qquad (12)$$

Expressing this estimate in terms of the original score scale,

$$n = \frac{(\mu_x - X_{min})(X_{max} - \mu_x) - r\sigma_x^2}{\sigma_x^2(1-r)} . \qquad (13)$$

Note the way in which the estimated effective test length depends on the possible score range. For a given distribution of scores with a given reliability coefficient, the larger the possible score range, the greater the estimated effective test length. Although the estimated effective test length is sensitive to this change in the specified possible score range, the estimated contingency tables are not. A large change in the specified minimum and maximum possible scores will produce very little change in the estimated contingency tables. The estimates of the conditional standard error of measurement, however, will change substantially, especially for true scores at the high and low ends of the score distribution. Some experimental results summarized later in this report illustrate the extent to which large changes in the specified minimum and maximum possible scores affected these statistics for one test.

Estimating the true-score distribution

The distribution of the proportional true scores (T_p) is estimated from the observed distribution of single-form scores (X), by a method developed by Lord (1965; also see Hanson, 1991, pp. 3-9). This method requires as input the first four moments of the distribution of the transformed score X'

(computed by transforming each individual observed score X to X'). Lord's (1965) method assumes that the proportional true score (T_p) has the form of a "four-parameter beta distribution", with density

$$\frac{1}{B(d+1,\Delta+1)} \frac{(T_p-a)^d (b-T_p)^\Delta}{(b-a)^{d+\Delta+1}} , \qquad (14)$$

where B is the beta function. This formula can be obtained by taking a random variable having a (two-parameter) beta distribution on (0,1), with parameters (d+1) and (Δ+1), and transforming it linearly onto the interval (a,b), where $0 \leq a < b \leq 1$. The additional parameters a and b make the model more flexible, by allowing zero frequency for extremely low or extremely high true-score levels.[1] One limitation of this model is that it does not allow for a bimodal true-score distribution. However, bimodal score distributions are fairly rare on tests of any substantial length. If the test score is a composite of two or more subscores measuring somewhat different proficiencies, a bimodal true-score distribution would be even less likely.

How good are the estimates?

The decision accuracy statistics estimated by this method describe the agreement between classifications based on an observable variable (scores on one form of a test) and classifications based on an unobservable variable (the test-takers' true scores). Therefore, these estimates cannot be evaluated on the basis of actual responses from real, live test-takers. In contrast, the

[1]Hanson and Brennan (1990) have shown that, when the estimated true-score distribution is used to reconstruct a single-form score distribution, the additional two parameters can greatly improve the fit of the reconstructed distribution to the observed distribution.

decision consistency statistics describe the agreement between classifications based on two observable variables (scores on two forms of the same test). These estimates can be evaluated in a situation in which the same test-takers take two alternate forms of a test, under conditions that make it reasonable to assume that test-takers' true scores will not change between forms. These conditions are not often found in actual test use. However, it is possible to create such a situation artificially, using actual test response data, by splitting a test into parallel halves. This split-halves approach was the basis for a series of tryouts of the method described in this paper. The procedure for the tryouts can be summarized as follows:

1. Select a test for which item scores are available for a large group of test-takers. Divide the test into two half-tests as similar as possible in the content and format of the items.
2. Compute the two half-test scores of each test-taker. Compute the distribution of scores on each half-test and the correlation between the half-test scores.
3. Select cut-points for each half-test score that represent, as closely as possible, the same percentile ranks as the cut-points actually used on the full test. At each cut-point, compute the 2x2 contingency table of classifications based on the two half-test scores. These tables indicate the "actual" consistency of the classifications based on the half-test scores.
4. Apply the estimation method to the scores on one of the half-tests, using the correlation between the two half-test scores as the reliability coefficient. Compare the resulting estimated 2x2 contingency tables with the "actual" tables from Step 3.

This evaluation procedure was applied to four tests. The tests were selected to be quite different from each other in content, format, and statistical characteristics. Test 1 was an all-multiple-choice test used in the licensing of elementary school teachers. States using this test use several different cut-points. The cut-points applied to the half-tests corresponded, in terms of percentile rank, to three of these user-state cut-points: the lowest, the highest, and one in the middle. The remaining three tests were from the Advanced Placement Program. They consisted of various combinations of multiple-choice items and constructed-response tasks of various types. The Advanced Placement "grade" reported to the test-taker and to the college is based on a weighted composite of scores on the different sections of the test. Four cut-points are applied to this composite score, to divide the composite score scale into five large intervals, representing the five Advanced Placement grades: 1 (lowest) to 5 (highest). The cut-points applied to the half-tests in this study were selected to represent, as nearly as possible (given the rounding of the composite scores), the same percentiles as the cut-points used on the full test.

Table 6 describes the half-tests used in the evaluation and presents the results. The descriptive information presented for each half-test includes the academic subject tested, the test-taker population and the number of test-takers for whom data were available, the number of items of each type,[2] the reliability coefficient, and the percentile ranks of the cut-points. The results are presented in terms of the proportion of the test-takers

---

[2]In two cases (English and Art History) the half-tests actually included less than half of the full Advanced Placement Examination. In each case the items left out were essays requiring approximately fifteen minutes of the test-taker's time.

consistently classified at each cut-point, i.e., classified in the same way by the two half-tests. The "actual result" is based on the agreement between the two half-tests. The "1st estimate" was obtained by applying the method described in this paper to the data from the first half-test. The "2nd estimate" was obtained by applying the method to the data from the second half-test.

For Test 1, all the estimates are within .005 of the actual values. For Test 2, all but one of the estimates are within .01 of the actual values, and four of the eight are extremely close. The one estimate that is not within .01 of the actual value misses by slightly more than .01. For Test 3, all the estimates are within .01 of the actual values. For Test 4, the estimates are within .01 of the actuals for cut-points 2 and 4, but they differ from the actuals by more than .02 for cut-points 1 and 3.

How much do the estimates depend on the possible score range?

To determine the extent to which the estimates produced by this method depend on the possible score range, we applied the method several times to the same data, changing the specified minimum and maximum possible scores. The data were from a test of 150 multiple-choice items, each scored 0 or 1, and 22 short-answer essay items, each scored 0 to 3. The total score was a weighted composite of scores on the two sections; the actual minimum and maximum possible scores were -20 and 223. When these values were input into the estimation procedure, the computer program did not run to completion, because the estimated effective test length was greater than it could accommodate. We then experimented by re-running the program five times, changing only the minimum and maximum possible scores. The widest of the five specified

possible score ranges was 1.76 times the observed score range; the narrowest of the five was identical to the range of scores actually observed.

Table 7 shows the results of the experiment. As the possible score range decreased, the estimated effective test length decreased greatly, from 238 items to 74 items. At the same time, the estimated percent correctly classified and the estimated percent consistently classified, at each of several cut-points, were almost completely unaffected by the change in the specified possible score range. Even where the effect was largest -- for cut-points near the median of the distribution -- the estimate changed by only about one-half of one percent. The estimated conditional standard error of measurement (CSEM), however, was sensitive to the change in the possible score range. (Note that the conditioning variable -- the true score -- was expressed on the scale of the scores themselves, rather than as a proportion of the possible score range.) As the possible score range narrows, score levels at the upper and lower ends of the distribution become closer to the maximum and minimum possible scores, and the estimated CSEM at these score levels decreases. (The CSEM for a test-taker whose true score is the maximum possible score or the minimum possible score must be zero.) Since the overall standard error of measurement does not change when the specified possible score range changes, the estimated CSEM in the middle of the distribution must increase, to compensate for the decrease at the ends of the distribution.

Conclusion

From the user's point of view, the essential features of the estimation method presented in this paper are these:

1. It estimates statistics describing the agreement between classifications based on alternate forms of a test (decision consistency) and between classifications based on one form and classifications based on test-takers' true scores (decision accuracy).

2. It requires as input only the distribution of scores on one form, the minimum and maximum possible scores, the cut-points used for classification, and the reliability coefficient. It will work for any test score for which this information is available, regardless of the format of the test.

3. The estimates of the percent of test-takers correctly classified and the percent of test-takers consistently classified tend to be within one percentage point of their actual values.

4. The estimates of the decision consistency and decision accuracy statistics are affected very little by large changes in the specified minimum and maximum possible scores.

References

Hanson, B. A. and Brennan, R. L., 1990. An investigation of classification consistency indexes estimated under alternative strong true score models. Journal of Educational Measurement, 27, 345-359.

Hanson, B. A., 1991. Method of moments estimates for the four-parameter beta compound binomial model and the calculation of classification consistency indexes. ACT Research Report Series, No. 91-5. Iowa City: ACT.

Huynh, H., 1976. On the reliability of domain-referenced testing. Journal of Educational Measurement, 13, 253-264.

Livingston, S. A. and Wingersky, M. S., 1979. Assessing the reliability of tests used to make pass/fail decisions. Journal of Educational Measurement, 16, 247-260.

Lord, F. M., 1965. A strong true-score theory, with applications. Psychometrika, 30, 239-270.

Subkoviak, M. J., 1976. Estimating reliability from a single administration of a criterion-referenced test. Journal of Educational Measurement, 13, 265-276.

Wilcox, R. R., 1981. A review of the beta-binomial model and its extensions. Journal of Educational Statistics, 6, 3-32.

Table 1.
Decision accuracy: estimated joint distribution (contingency table)

Cell entry is proportion of all examinees

Classification on form taken	Classification on all-forms average*					Total
	Category A	Category B	Category C	Category D	Category E	
Category A (161 to 200)	.0248	.0148	.0002	.0000	.0000	.0398
Category B (143 to 160)	.0079	.0746	.0277	.0000	.0000	.1102
Category C (103 to 142)	.0001	.0236	.3581	.0359	.0000	.4176
Category D (60 to 102)	.0000	.0000	.0339	.2748	.0127	.3213
Category E (0 to 59)	.0000	.0000	.0000	.0193	.0918	.1111
Total	.0328	.1130	.4198	.3299	.1045	1.0000

Actual number of examinees 1080

* "True score"

Table 2.
Decision accuracy: estimated 2x2 contingency tables.

Cell entry is proportion of all examinees

Classification on form taken		Classification on all-forms average* Category A	Categories B,C,D,E
Examinees in Category A	(161+)	.025	.015
Examinees in Categories B,C,D,E	(160-)	.008	.952
Estimated proportion correctly classified =		.977	

		Categories A,B	Categories C,D,E
Examinees in Categories A,B	(143+)	.122	.028
Examinees in Categories C,D,E	(142-)	.024	.826
Estimated proportion correctly classified =		.948	

		Categories A,B,C	Categories D,E
Examinees in Categories A,B,C	(103+)	.532	.036
Examinees in Categories D,E	(102-)	.034	.399
Estimated proportion correctly classified =		.930**	

		Categories A,B,C,D	Category E
Examinees in Categories A,B,C,D	(60+)	.876	.013
Examinees in Category E	(59-)	.019	.092
Estimated proportion correctly classified =		.968	

Actual number of examinees 1080

* "True score "
** Inconsistent with cell entries because of rounding

Table 3.
Decision consistency: estimated joint distribution (contingency table)

Cell entry is proportion of all examinees

Classification on form taken	Classification on alternate form: Category A	Category B	Category C	Category D	Category E	Total
Category A (161 to 200)	.0237	.0150	.0011	.0000	.0000	.0398
Category B (143 to 160)	.0155	.0616	.0331	.0000	.0000	.1102
Category C (103 to 142)	.0013	.0375	.3269	.0518	.0000	.4176
Category D (60 to 102)	.0000	.0000	.0463	.2531	.0219	.3213
Category E (0 to 59)	.0000	.0000	.0000	.0230	.0881	.1111
Total	.0405	.1141	.4075	.3280	.1099	1.0000

Actual number of examinees 1080

Table 4.
Decision consistency: estimated 2x2 contingency tables.

Cell entry is proportion of all examinees

Classification on form taken		Classification on alternate form: Category A	Categories B,C,D,E
Examinees in Category A	(161+)	.024	.016
Examinees in Categories B,C,D,E	(160-)	.017	.943
Estimated proportion consistently classified =		.967	

		Categories A,B	Categories C,D,E
Examinees in Categories A,B	(143+)	.116	.034
Examinees in Categories C,D,E	(142-)	.039	.811
Estimated proportion consistently classified =		.927	

		Categories A,B,C	Categories D,E
Examinees in Categories A,B,C	(103+)	.516	.052
Examinees in Categories D,E	(102-)	.046	.386
Estimated proportion consistently classified =		.902	

		Categories A,B,C,D	Category E
Examinees in Categories A,B,C,D	(60+)	.867	.022
Examinees in Category E	(59-)	.023	.088
Estimated proportion consistently classified =		.955	

Actual number of examinees 1080

Table 5.
Score scales used in the estimation process.

Name	Symbol	Range	Size of interval
Single-form score	X	X_{min} to X_{max}	1
	(X_0, X_1, X_2)		
Proportional single-form score	p	.00 to 1.00	.01
Transformed single-form score	X'	0 to n	1
Proportional true score	T_p	.00 to 1.00	.01

Table 6.
Results of the evaluation of the method.

Note: The statistics in this table describe the half-tests used in the evaluation, not the full parent tests.

Test 1

Subject: All elementary school subjects
Test-taker population: Beginning elementary school teachers and student teachers
Number of test-takers: 10,352
Format (half-test): 74 multiple-choice items
Reliability coefficient: .852

Percentile rank of each cut-point:

	Cut 1	Cut 2	Cut 3
1st half-test	13.5	8.1	2.9
2nd half-test	14.3	8.7	2.9

Proportion of test-takers classified consistently at each cut-point:

	Cut 1	Cut 2	Cut 3
Actual	.923	.945	.976
1st estimate	.919	.947	.976
2nd estimate	.919	.947	.978

Test 2

Subject: English literature and composition
Test-taker population: High school students seeking advanced placement
Number of test-takers: 113,129
Format (half-test): 23 multiple-choice items and one essay, scored 0 to 9
Reliability coefficient: .603

Percentile rank of each cut-point:

	Cut 1	Cut 2	Cut 3	Cut 4
1st half-test	87.2	69.5	31.1	2.4
2nd half-test	87.6	69.0	31.1	2.6

Proportion of test-takers classified consistently at each cut-point:

	Cut 1	Cut 2	Cut 3	Cut 4
Actual	.850	.733	.749	.966
1st estimate	.853	.744	.749	.965
2nd estimate	.849	.739	.751	.962

Table 6 (continued).
Results of the evaluation of the method.

Test 3

Subject: French language
Test-taker population: High school students seeking advanced placement
Number of test-takers: 3,555
Format (half-test): 36 multiple-choice items, 10 completion items, and either one speaking task, scored 0 to 9, or 6 speaking questions, scored 0 to 4.
Reliability coefficient: .891

Percentile rank of each cut-point:

	Cut 1	Cut 2	Cut 3	Cut 4
1st half-test	84.6	67.6	33.2	15.7
2nd half-test	84.8	68.6	34.0	15.8

Proportion of test-takers classified consistently at each cut-point:

	Cut 1	Cut 2	Cut 3	Cut 4
Actual	.914	.870	.871	.909
1st estimate	.917	.872	.862	.902
2nd estimate	.905	.866	.867	.912

Test 4

Subject: Art history
Test-taker population: High school students seeking advanced placement
Number of test-takers: 1,960
Format (half-test): 53 multiple-choice items and three short essays, scored 1 to 4.
Reliability coefficient: .641

Percentile rank of each cut-point:

	Cut 1	Cut 2	Cut 3	Cut 4
1st half-test	86.3	65.4	27.9	9.1
2nd half-test	86.5	65.0	28.1	9.5

Proportion of test-takers classified consistently at each cut-point:

	Cut 1	Cut 2	Cut 3	Cut 4
Actual	.833	.753	.785	.884
1st estimate	.855	.748	.762	.881
2nd estimate	.856	.745	.763	.881

Table 7.
Results of varying the maximum and minimum possible scores.

(Observed score range = 52 to 183; mean score = 119.2;
standard deviation = 23.8; reliability = .91)

Maximum possible score		210	200	194	188	183
Minimum possible score		-20	1	24	47	52
Possible range/ observed range		1.76	1.52	1.30	1.08	1.00
Estimated effective test length		238	178	130	87	74
Estimated percent correctly classified :						
Cut-point	percentile					
151.5	90.0	.959	.959	.959	.958	958
136.5	75.0	.931	.931	.930	.927	.927
118.5	49.7	.909	.909	.907	.905	.905
102.5	25.6	.920	.921	.921	.921	.922
88.5	10.4	.951	.952	.953	.955	.956
79.5	4.8	.969	.970	.971	.974	.974
69.5	1.0	.991	.991	.991	.992	.992
Estimated percent consistently classified:						
Cut-point	percentile					
151.5	90.0	.942	.943	.943	.942	.941
136.5	75.0	.902	.902	.901	.897	.897
118.5	49.7	.873	.872	.871	.868	.868
102.5	25.6	.889	.888	.889	.889	.890
88.5	10.4	.931	.931	.933	.936	.937
79.5	4.8	.958	.958	.960	.963	.964
69.5	1.0	.985	.985	.986	.988	.988
Estimated conditional standard error of measurement:						
True score*	percentile					
180	99.9	5.04	4.51	4.13	3.53	2.26
160	95.3	6.16	5.99	5.98	6.06	5.82
140	78.3	6.87	6.86	6.96	7.18	7.17
120	51.1	7.29	7.33	7.41	7.57	7.63
100	21.5	7.46	7.47	7.43	7.34	7.36
80	4.8	7.40	7.31	7.02	6.42	6.27
60	0.3	7.11	6.83	6.11	4.41	3.68

*Expressed in the units of the scores